A Galaxy-based, multi-staged, two-step searching pipeline for improved peptide identification in next generation proteomic studies

Praveen Kumar[1], James Johnson[2], Matthew C. Chambers[3], Mohammad Heydarian[4], Thomas McGowan[2], Joel D. Rudney[5], Pratik Jagtap[6], Timothy Griffin[6]

[1]Biomedical Informatics and Computational Biology, University of Minnesota, Minneapolis, MN; [2]Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, MN; [3]Department of Biochemistry, Vanderbilt University, Nashville, TN; [4]Department of Biology, Johns Hopkins University, Baltimore, MD; [5]Department of Diagnostic and Biological Sciences, School of Dentistry, University of Minnesota, Minneapolis, MN; [6]Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, Minneapolis, MN;

Introduction

In a proteogenomic study (Figure 1), genome/transcriptome sequencing data is used for the generation of potentially expressed protein variants. This produces a large number of possible sequences to be matched with (MS/MS) data.

Matching the MS/MS data to peptide sequences contained within a database confirms the presence of translated variant proteins in the sample.

One of the challenge proteogenomics analysis faces is the large databases. It increases false-positive identifications, leading to loss of sensitivity for identifying true peptide spectrum matches (PSMs).

Similarly, a metaproteomics study encounters the challenge of large databases as the database being used contains protein sequences from thousands of microbes.

We developed the multi-stage database searching (multiDBsearch) method (Figure 2) that can address these challenges, and it may also help in optimizing the database composition being used for matching.

We have implemented multiDBsearch method in the flexible and user-friendly Galaxy platform (Figure 3), where we have evaluated its utility for proteogenomics and metaproteomics applications.

Figure 1: Proteogenomics Pipeline

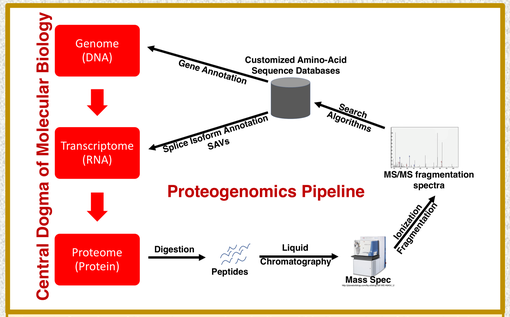

Figure 1: Overview of the proteogenomics pipeline within the context of the central dogma of molecular biology

Figure 2: MultiDBsearch (Multi-stage database searching)

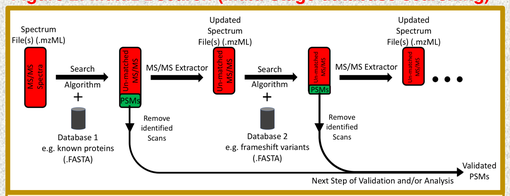

Figure 2: MultiDBsearch (Pipeline describing the matching of MS/MS data against a database, removing the matched spectra, and matching the un-matched MS/MS spectra against next database)

Workflow and Method Evaluation

Figure 3: Galaxy Workflow

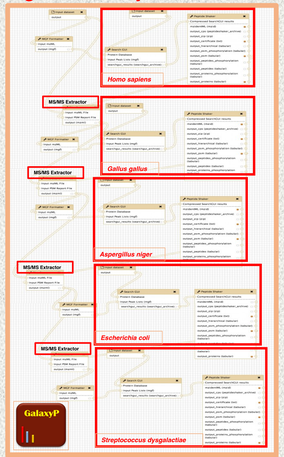

Figure 3: Galaxy Workflow for multiDBsearch

Figure 4: 2-Step Search Method

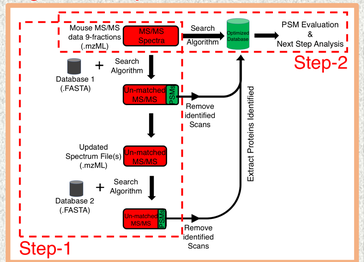

Figure 4: Two-Step searching method after multiDBsearch

Method	# of Database	Database searching method				
1 Ideal case	1	Human proteome + contaminants + 4 non-human proteins				
2 Traditional Method	1	Human proteome + contaminants + 4 non-human organisms' proteome				
3 MultiDBsearch	5	Human + contaminants	*G. gallus*	*A. niger*	*E. coli*	*S. dysgalactiae*
4 Two-step method	1	Human proteome + contaminants + proteins identified from 4 non-human organisms (from method 3)				
5 Two-step method	1	Proteins identified from all 5 organisms (from method 3)				

Table 1: Table 1: MS/MS data matching methods and databases used for evaluation

ABRF Data

- MS/MS Data: Human proteins + 4 spiked-in proteins:
  1) Beta-Gal from *Escherichia Coli*
  2) Lysozyme from *Gallus gallus*
  3) Protein-G from *Aspergillus niger*
  4) Amylase from *Streptococcus dysgalactiae*
- MS/MS data matched with databases in 5 different methods (Table 1)
- Ideally, we should identify only the spiked-in proteins when MS/MS data is matched with those organisms' proteome.
- Tested two-step search method (Figure 4), where a smaller database is created based on the multiDBsearch protein identifications.

Evaluation Results

- Using multiDBsearch method, more number of human proteins were identified and less number non-human proteins identified (Figure 5a-b)
- Less peptides and PSMs identified for spiked-in proteins *(Figure 5c-d)*
- Adding two-step method to multiDBsearch method helps in rescuing

Figure 5: Protein & PSM Numbers

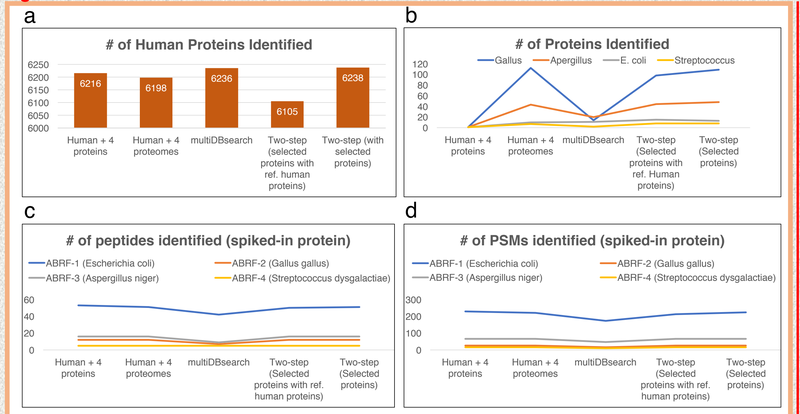

Figure 4: Results obtained from different approaches for ABRF standard data (a) Number of human proteins identified (b) Number of proteins identified from non-human organisms (c) Number of peptides identified for spiked-in protein, and (d) Number of PSM identifications for spiked-in protein.

Proteogenomics

Mouse MS/MS Data + RNA-Seq Data

- Protein sample from mouse pre-pro-B cells and pro-B cells
- MS/MS data searched against following databases (Table 2):
  1) *Mouse Uniprot protein sequences*
  2) *3-frame translated cDNA (from EMBL)*
  3) *Splice junction variants (derived from RNA-Seq data)*
  4) *3-frame translated long non-coding RNA*
  5) *Single amino-acid variants (partial database)*
- MS/MS data searched in three ways using:
  1) *All databases combined (single search)*
  2) *MultiDBsearch*
  3) *2-Step search with database derived from multiDBsearch*

Method	Number of entries in fasta DB			Number of Novel PSMs		
	All DB combined	multiDBsearch	2-Step Database	All DB combined	multiDBs earch	2-Step Database
UniProt mouse + Contaminants	25,091	25,091	17,925	NA	NA	NA
3-frame translated cDNA	2,087,787	2,087,787	39,924	125	192	573
Splice junction variants	68,122	68,122	16,157	0	5	13
Long non-coding RNA	420,936	420,936	22514	34	1	116
Single amino acid variations	77	77	77	6	7	2
Total	2,602,013	NA	96597			

Table 2: Database size (number of fasta sequence entry) used in each method for proteogenomic analysis and the number of novel PSMs identified by each method

Figure 6: Post-processing Validation Steps

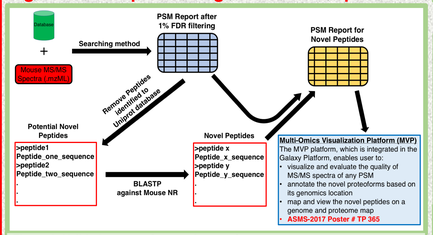

Figure 6: Validation steps for novel proteoforms

Figure 7: Identified PSM Numbers

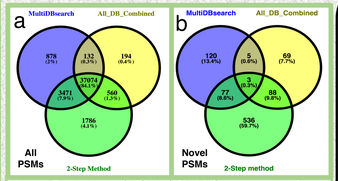

Figure 7: Number of PSMs (1% FDR) compared between three methods (a) from all databases (b) only novel peptides after blastp step (Figure 6)

Results

- Two-Step search method is helping in identifying more PSMs (Figure 7a-b)
- Two-Step search method allowing to identify more number of novel proteoforms
- MultiDBsearch helps optimize the database (Table 2)
- Search against optimized DB helps identifying more PSMs

Metaproteomics

Oral Plaque Metaproteome Data

- Protein sample from oral plaque grown with sucrose
- Database used: Human Oral Microbiome Database (HOMD)
- Size of HOMD: 1,079,626 (method 1: searched all combined)
- Divided into 5 sections of ~200,000 (used multiDBsearch)
- Two-Step searching after multiDBsearch
  - Helped optimizing the database (210,425 sequences)
  - 10% More PSMs identified which can help getting better depth in Unipept analysis (Figure 9) and functional analysis

Figure 8: Metaproteomics by sectioning DB

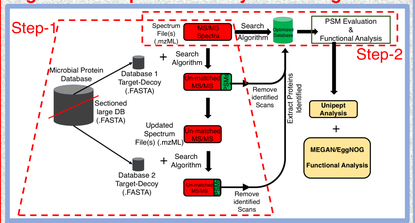

Figure 8: Metaproteomic data analysis pipeline. Two-step searching after sectioning the database and using multiDBsearch.

Figure 9: Unipept Results

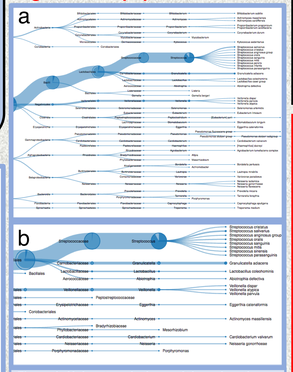

Figure 9: Unipept analysis result based on the peptides identified (a) Two-Step searching method (b) all HOMD searched together

#genus	Streptococcus	Veillonella	Granulicatella	Lactobacillus	Propionibacterium	Eggerthia	Neisseria	Bifidobacterium	Actinomyces	Prevotella	Pseudomonas	Tannerella	Variovorax
2-Step search PSMs	359	36	33	11	7	3	3	3	2	2	2	2	2
All combined search PSMs	352	31	34	10	0	2	2	0	1	0	0	0	0

Table 3: Number of PSMs identified for each genus after Unipept analysis (Table shows only those genera for which at least 2 PSMs were identified)

Future Directions

- Setting a cutoff based on the precursor mass error distribution to select proteins in optimizing the database
- Validation of PSMs using MVP platform and characterizing the novel proteoforms
- Functional analysis on metaproteomic results using MEGAN/EggNOG
- Oral plaque data is available for both grown with sucrose and without sucrose
  - Carry out comparative analysis using Two-step searching method
  - Use 16S-rRNA data to create database and carry out analysis

Acknowledgements

- Data used for evaluation was generated through the collaborative work of the ABRF Proteomics Research Group.
- This project is supported by National Science Foundation (NSF) grant "1458524" and National Institutes of Health (NIH) grant "U24CA199347"